The relationship of Yellow card and Full-Time Away Team Goals'*

Analysis of Primer league 2023-24 season

Che-Yu Wang

April 18, 2024

Football match outcomes, especially Full-Time Away Team Goals (FTAG), are difficult to predict in sports analytics. This study uses 2024 Premier League data to examine how Home Team Yellow Cards (HY) affect away team scoring. The study uses logistic, Poisson, and negative binomial regressions to find the best statistical model for football score count data. Poisson regression, ideal for count data, assumes equal mean and variance, which is often violated in real-world data. Logistic regression models binary outcomes and may overlook nuanced count data. Negative binomial regression better represents data variability by adding an overdispersion parameter. The study evaluates each model's ability to capture football scoring patterns' complexities using statistical analysis, model comparison, and diagnostic measures. The findings emphasize the delicate balance between model complexity and predictive accuracy, emphasizing the importance of choosing the right model to understand away team goal-scoring trends. This meticulous approach improves football analytics predictive modelling and sports statistics by refining count-based outcome analysis..

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset aims to provide a comprehensive analysis of football match outcomes, focusing on factors influencing match results like goals scored by the away team and yellow card. It also aims to develop and validate predictive models for forecasting match outcomes based on historical data. The dataset serves as a valuable resource for researchers, analysts, and enthusiasts interested in understanding the dynamics

*Code and data are available at: LINK https://github.com/Cheyuwang/Primer-league-analysis.

of football matches and exploring trends over time. It fills a critical gap by offering structured and detailed information for thorough analyses and decision-making in football analytics. .

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - football data UK
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - football data UK
4. *Any other comments?*
   - no

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The dataset includes instances of individual football games, each with information on the performance of the home team, goals scored by the opposition, and other match-specific details. Due to its homogeneous collection of cases, match results and associated variables can be examined and analyzed in depth. Every observation relates to a unique football-related event, avoiding various kinds of occurrences or intricate relational frameworks.
2. *How many instances are there in total (of each type, if appropriate)?*
   - around 100
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset is a compilation of football match records; it may include every Premier League match played in 2024. The comprehensiveness, selection process, and inclusion of different teams, seasons, and geographical areas all contribute to the data's representativeness. Its representativeness must be confirmed by contrasting sample statistics with the known demographics of the larger group. If the dataset is

not representative, it may be because of restricted access to the data or a purposeful focus on a particular subset. It is essential to validate or acknowledge bias if one hopes to maintain the integrity of later analyses.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - The Premier League matches dataset for the 2024 season includes match dates, unique match IDs, team names, team statistics, player-specific data, match officials, play location, and key match outcomes. This data provides detailed insights into individual matches, allowing for precise separation of events and serving a variety of statistical, tactical, and predictive modeling purposes in sports analytics. The dataset also includes time-series data, which provides a minute-by-minute breakdown of match events, resulting in a dynamic narrative of play-by-play developments.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - Each instance in the Premier League dataset for the 2024 season is likely to include a target or label that serves as a focus for analysis, such as predicting the match outcome (win, loss, or draw for the home team), forecasting the number of goals scored, or investigating the frequency and impact of disciplinary actions (yellow/red cards). If the dataset's goal is to investigate the effects of home team yellow cards on away team success, the target variable for each instance could be the number of goals scored by the away team during the game. This allows for a clear analysis of whether there is a correlation between disciplinary measures indicated by yellow cards and the away team's scoring performance.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - No

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Football datasets, such as the Premier League 2024 season, rarely include relationships between match instances because each game is considered an independent event. Implicit relationships include temporal sequences, in which previous matches may influence future games, and team-related connections, in which matches involving the same team throughout the season share strategies and player dynamics. League standings are also influenced by matches within the same season, which

are played under uniform rules and competitive conditions. Data must be manipulated to reveal these inherent but unstated relationships in order to conduct more in-depth analysis, such as tracking match results or team performance.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No, the data already professional and clean

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - No

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - It's critical to determine whether the Premier League 2024 dataset is self-contained or depends on external resources like websites, player databases, or social media links. If the dataset includes external data, concerns about the long-term availability and consistency of these resources arise, as websites change or are removed over time. There may not be official archival versions of the dataset that include these external elements, which could limit the reproducibility of research or analysis based on this data. Furthermore, external resources may be subject to licenses, fees, or other restrictions that limit how the dataset is used or accessed by others. Users of datasets should thoroughly review any external dependencies and their conditions in order to fully understand the dataset's utility as well as any limitations or costs associated with its use.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Individuals are highly likely to be identified in the Premier League 2024 dataset, as sports datasets typically contain detailed information about players, coaches, and possibly referees. This information may include names, positions, statistical performance data, and possibly images or jersey numbers, all of which are typically intended for public dissemination and fan engagement. Explicit identifiers such as names and jersey numbers allow for direct identification. Furthermore, when combined with other publicly available data, such as news articles, social media profiles, or historical performance data, the possibility of indirect identification grows, allowing for a more complete profile of these individuals. Thus, while identification is inherent in the nature of sports datasets, there is an obligation to handle any sensitive personal data responsibly and in accordance with data protection regulations.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - For accuracy and reliability, official sports statisticians and automated systems like goal-line technology and VAR observe most matches in the Premier League 2024 dataset. Match results, player statistics, and team performances are usually

recorded live and publicly available. Player ratings and predictive metrics are derived from observed data using statistical models and validated against benchmarks or historical data for reliability. Subject-reported data would require additional validation checks like consistency with known information or cross-referencing with medical records, but this is rare in observational sports datasets.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The Premier League 2024 dataset is gathered using both hardware and software. Cameras, goal-line technology, and other sensor-based systems monitor and record live match events, which are then aggregated and organized by APIs and sports analytics platforms. Human curation, specifically data entry verification and correction, ensures statistical accuracy. To meet sports data analytics standards, these collection mechanisms are validated through hardware calibration and software algorithm updates. Quality assurance processes ensure that data collection accuracy is consistent and reliable across games and seasons.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - To guarantee the representativeness and dependability of the data, it's critical to comprehend the sampling strategy used if the Premier League 2024 dataset is a sample from a larger set. Sampling strategies in sports analytics can be deterministic, wherein events or matches are selected based on predetermined criteria (e.g., only derby matches or high stakes games), or probabilistic, wherein matches are selected based on predefined probabilities that represent a variety of scenarios or team distributions. The approach selected affects the analysis and findings' generalizability. Examining the plan for bias and coverage is usually part of validating the sampling technique. This involves making sure that the sample fairly represents the larger set in terms of team performance, game results, and other crucial elements influencing the dynamics of the season.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Player and referees in Primer League

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- Data collection for a dataset such as Premier League 2024 would normally take place between August 2023 and May 2024, which is the duration of the 2023–2024 football season. Given that the data represents the actual events and results of the football matches, this timeline would exactly align with the timeframe during which the data associated with each instance was created. Every event in the dataset, including team performances, player statistics, and match results, is captured as it happens, guaranteeing that the data is up to date with the events it chronicles. The Premier League data is current and directly reflects the season as it unfolds, in contrast to retrospective data collections like crawls of old news articles, which have a lag between the occurrence of the matches and their documentation in the dataset.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?* The Premier League 2024 dataset's data was gathered from third-party sources, including club websites, official Premier League match statistics, and sports analytics services, rather than directly from the players. These resources gather and distribute player statistics, team performance metrics, and match data; they are all freely available to the public and frequently utilized in sports reporting and analytics. This approach to data collection guarantees that the information is complete and consistent, following the conventions and formats common to the collection of sports data. Furthermore, third-party sources frequently offer improved accuracy and dependability thanks to expert data management techniques, such as real-time data updates and verification during sporting events.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - No

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - The question about the data were necessary

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - In the case of the Premier League 2024 dataset, which is mainly made up of publicly accessible sports performance data, players' or teams' individual consent is typically not needed because the data is public and not private. Because the data is gathered and used in the public domain and in accordance with professional sports rules and agreements, there is usually no way for these individuals to withdraw their consent for its use.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No, I didn't

12. *Any other comments?*

    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - I used R packages, including dplyr, readr, model summary, janitor, tibble, and ggplot2, were used for data preprocessing, cleaning, and labeling. The study mentions data manipulation and reading from external sources but does not go into specific pre-processing steps. But the document makes no mention of particular pre-processing methods for compiling and natural language processing.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - No

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No

4. *Any other comments?*

- No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - A Poisson regression analysis has already been performed using the Premier League 2024 dataset to investigate the correlation between the number of yellow cards the home team receives and the goals scored by the away team at full time. This statistical method aids in determining whether an increase in goals scored by rival teams is correlated with a higher frequency of disciplinary actions against the home team. For both sports analysts and team strategists, the analysis can offer valuable insights into how team behavior may affect match outcomes.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - https://github.com/Cheyuwang/Primer-league-analysis.
3. *What (other) tasks could the dataset be used for?*
    - No
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - The Premier League 2024 dataset is primarily composed of objective performance data. However, biases may occur if certain teams or player groups are disproportionately highlighted or overlooked during data collection or preprocessing. This could result in skewed analytics or unfair representations. In order to reduce these risks, users of datasets should make sure that data is used in a diverse and balanced manner, confirm findings with other data sources, and be aware of the dataset's boundaries and scope to prevent erroneous interpretations that might affect public opinion or policy decisions.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - no
6. *Any other comments?*
    - no

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, The document describes an open access dataset that can be found in a public GitHub repository. It is possible for third parties other than the organization that created the dataset to receive access to the Premier League 2024 dataset. Sports analytics datasets are frequently shared with collaborators, scholarly researchers, and for-profit companies that are interested in sports performance analysis and statistics. Usually, these distributions are controlled by license agreements that guarantee the data is used morally and compliantly with privacy laws. Third parties must strictly abide by these terms in order to prevent misuse and to encourage transparency and responsible data handling practices.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is available on GitHub, a platform for code sharing and collaboration. This tool allows for version control, issue tracking, and user discussions. The document does not specify if the dataset has a Digital Object Identifier (DOI), but users can learn more about DOIs from GitHub repositories or academic publications3.

3. *When will the dataset be distributed?*

    - Since the study's release, the dataset has been made available to the public on GitHub. Users can access the dataset by clicking on the provided link, but since open-source projects are dynamic, it's crucial to regularly check the repository for updates and changes. Examining the release section or commit history can provide insight into the evolution of the project.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset's copyright or intellectual property license, along with any usage restrictions or costs, are not specified in the document. But for academic and research datasets, the open Creative Commons licenses—which permit different levels of reuse and redistribution—are frequently utilized. Users who would like to review licensing details, including attribution and restrictions, should go to the appropriate 9 GitHub repository. To ensure ethical and legal use, they are advised to check the dataset's GitHub repository for the most recent licensing terms.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other*

access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

- No

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - The document does not specify the dataset's copyright, license, terms of use, or fees. Academic and research datasets often use Open Creative Commons licenses, which allow reuse and redistribution. Users should check the relevant GitHub repository for licensing information like attribution and restrictions. They should check the dataset's GitHub repository for the latest licensing terms to ensure ethical and legal use.

7. *Any other comments?*

   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The Premier League's official statistics department and football data uk or a sports analytics company may host the 2024 Premier League dataset. These entities keep the dataset current, accurate, and user-friendly. In addition to answering user questions, they verify, quality control, and update the dataset throughout the season.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - It does not provide

3. *Is there an erratum? If so, please provide a link or other access point.*

   - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - The document makes no mention of changing communication strategies or providing support for older versions of the dataset. Older versions can be accessed on GitHub via releases and commit histories, but dataset creators are still responsible for updating and maintaining them. If support is stopped, messages can be sent via README files, release notes, or issue discussions. The repository should be frequently checked by users for updates.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - No, but GitHub allows for the archival and access of older versions via commit history and releases, but dataset creators must actively support or update them

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - No

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - No

8. *Any other comments?*

    - No

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.